CLAIMS

1. A method for identifying a set of genes from a multiplicity of genes whose expression levels at a first state and a second state are measured in replicates using one or more nucleotide arrays, thereby generating a first plurality of independent measurements of the expression levels for said first state and a second plurality of independent measurements of the expression levels for said second state, which method comprises the following sequential steps:

(a) identifying a quality function capable of evaluating the distinctiveness between the first plurality and the second plurality;

(b) forming a first predetermined number of permutations from the first and the second pluralities, dividing said permutations into a first permutated plurality and a second permutated plurality, corresponding in size, to said first and second plurality, respectively, and identifying groups of genes the size of which is a second predetermined number, wherein the values of the quality function for the group of genes in said first permutated and second permutated pluralities attain the maximum;

(c) determining, from said first and second permutated pluralities, the top α^{th} percentile of the null distribution based on a quantitative characteristic of said groups of genes;

(d) identifying, based on the first and second pluralities, a subset of genes the size of which is said second predetermined number, wherein the values of the quality function for said subset of genes in said first and second pluralities attain the maximum;

(e) adding to the set of genes, said subset, if the value of said quantitative characteristic associated with said subset exceeds said top α^{th} percentile of the null distribution; and

(f) removing from the first and second pluralities, all measurements on said subset, if the maximum value of the quality function associated with said subset exceeds said top α^{th} percentile of the null distribution, and repeating steps (d)-(f) until no more measurements are left in the first and second pluralities or the value of said quantitative characteristic associated with the subset does not exceed said top α^{th} percentile of the null distribution.

2. The method of claim 1, wherein said states are selected from the group consisting of biological states, physiological states, pathological states, and prognostic states.

3. A method for identifying a set of genes from a multiplicity of genes whose expression levels in a first tissue and a second tissue are measured in replicates using one or more nucleotide arrays, thereby generating a first plurality of independent measurements of the expression levels for said first tissue and a second plurality of independent measurements of the expression levels for said second tissue, which method comprises:

(a) identifying a quality function capable of evaluating the distinctiveness between the first plurality and the second plurality;

(b) forming a first predetermined number of permutations from the first and the second pluralities, dividing said permutations into a first permutated plurality and a second permutated plurality, corresponding in size to said first and second plurality, respectively, and identifying groups of genes the size of which is a second predetermined number, wherein the values of the quality function for the group of genes in said first permutated and second permutated pluralities attain the maximum;

(c) determining, from said first and second permutated pluralities, the top α^{th} percentile of the null distribution based on a quantitative characteristic of said groups of genes;

(d) identifying, based on the first and second pluralities, a subset of genes the size of which is said second predetermined number, wherein the values of the quality function for said subset of genes in said first and second pluralities attain the maximum;

(e) adding to the set of genes, said subset, if the value of said quantitative characteristic associated with said subset exceeds said top α^{th} percentile of the null distribution; and

(f) removing from the first and second pluralities, all measurements on said subset, if the maximum value of the quality function associated with said subset exceeds said top α^{th} percentile of the null distribution, and repeating steps (d)-(f) until no more measurements are left in the first and second pluralities or the value of said quantitative characteristic associated with the subset does not exceed said top α^{th} percentile of the null distribution.

4. The method of claim 3, wherein said tissues are selected from the group consisting of normal lung tissues, cancer lung tissues, normal heart tissues, pathological heart tissues, normal and abnormal colon tissues, normal and abnormal renal tissues, normal and abnormal prostate tissues, and normal and abnormal breast tissues.

5. A method for identifying a set of genes from a multiplicity of genes whose expression levels in a first type of cells and a second type of cells are measured in replicates using one or more nucleotide arrays, thereby generating a first plurality of independent measurements of the expression levels for said first type of cells and a second plurality of independent measurements of the expression levels for said second types of cells, which method comprises:

(a) identifying a quality function capable of evaluating the distinctiveness between the first plurality and the second plurality;

(b) forming a first predetermined number of permutations from the first and the second pluralities, dividing said permutations into a first permutated plurality and a second permutated plurality, corresponding in size, to said first and second plurality, respectively, and identifying groups of genes the size of which is a second predetermined number, wherein the values of the quality function for the group of genes in said first permutated and second permutated pluralities attain the maximum;

(c) determining, from said first and second permutated pluralities, the top α^{th} percentile of the null distribution based on a quantitative characteristic of said groups of genes;

(d) identifying, based on the first and second pluralities, a subset of genes the size of which is said second predetermined number, wherein the values of the quality function for said subset of genes in said first and second pluralities attain the maximum;

(e) adding to the set of genes, said subset, if the value of said quantitative characteristic associated with said subset exceeds said top α^{th} percentile of the null distribution; and

(f) removing from the first and second pluralities, all measurements on said subset, if the maximum value of the quality function associated with said subset exceeds said top α^{th} percentile of the null distribution, and repeating steps (d)-(f) until no more measurements are left in the first and second pluralities or the value of said quantitative characteristic associated with the subset does not exceed said top α^{th} percentile of the null distribution.

6. The method of claim 5, wherein said types of cells are selected from the group consisting of normal lung cells, cancer lung cells, normal heart cells, pathological heart cells, normal and abnormal colon cells, normal and abnormal renal cells, normal and abnormal prostate cells, and normal and abnormal breast cells.

7. The method of claim 5, wherein said type of cells are selected from the group consisting of cultured cells and cells isolated from an organism.

8. The method of claim 1, 3, or 5, wherein said quality function is a probability distance function.

9. The method of claim 8, wherein said probability distance function is selected from the group consisting of the Mahalanobis distance and the Bhattacharya distance.

10. The method of claim 8, wherein the probability distance function is defined as:

$$N(\mu,\nu) = 2\int_{R^d}\int_{R^d} L(x,y)d\mu(x)d\nu(y) - \int_{R^d}\int_{R^d} L(x,y)d\mu(x)d\mu(y) - \int_{R^d}\int_{R^d} L(x,y)d\nu(x)d\nu(y)$$

where μ and ν are two probability measures defined on the Euclidean space, and $L(x,y)$ is a strictly negative definite kernel.

11. The method of claim 10, wherein the negative definite kernel is combined with the Euclidean distance between x and y to form a composite kernel function.

12. The method of claim 1, 3, or 5, wherein the quantitative characteristic is selected from the group consisting of an associated probability distance, a test set classification rate, and a cross validation classification rate.

13. The method of claim 1, 3, or 5, wherein the formation of the permutations further comprises:

(i) shifting the measurements in the first and second pluralities such that the marginal means thereof share the same true mean; and

(ii) randomly permuting the resulting shifted measurements thereby forming a null-distribution of permutations.

14. The method of claim 1, 3, or 5, wherein the identifying further comprises:

(i) calculating the values of the quality function for said subset of genes in said first and second pluralities thereby evaluating the distinctiveness of said first and second pluralities;

(ii) substituting a gene in said subset with one outside of said subset, thereby generating a new subset, and repeating step (i), keeping the new subset if the distinctiveness increases and the original subset if otherwise; and

(iii) repeating steps (i) and (ii) for a fourth predetermined number of times.

15. The method of claim 1, 3, or 5, wherein the identifying further comprises:

(i) randomly dividing the first and the second pluralities into v groups of an approximate equal size;

(ii) removing one of said v groups from said first and second pluralities and identifying, from the resulting reduced first and second pluralities, a subset of genes for which the value of said quality function attains the maximum; and

(iii) repeating step (ii) for each of said v groups thereby obtaining v subsets of genes.

16. The method of claim 1, 3, or 5, wherein the nucleotide arrays are selected from the group consisting of arrays having spotted thereon cDNA sequences and arrays having synthesized thereon oligonucleotides.